数据挖掘的 10 大分析方法

不仅仅是选中的十大算法，其实参加评选的 18 种算法，实际上随便拿出一种来都可以称得上是经典算法，它们在数据挖掘领域都产生了极为深远的影响。

C4.5 算法是机器学习算法中的一种分类决策树算法,其核心算法是 ID3 算法.C4.5 算法继承了 ID3 算法的优点，并在以下几方面对 ID3 算法进行了改进：

1)用信息增益率来选择属性，克服了用信息增益选择属性时偏向选择取值多的属性的缺乏；

2)在树构造过程中进行剪枝；

3)能够完成对连续属性的离散化处理；

4)能够对不完整数据进行处理。

C4.5 算法有如下优点：产生的分类规那么易于理解，准确率较高。其缺点是：在构造树的过程中，需要对数据集进行屡次的顺序扫描和排序，因而导致算法的低效。

2.Thek-meansalgorithm 即 K-Means 算法

k-meansalgorithm 算法是一个聚类算法，把 n 的对象根据他们的属性分为 k 个分割，
k 3.Supportvectormachines

支持向量机，英文为 SupportVectorMachine，简称 SV 机〔论文中一般简称 SVM〕。它是一种监督式学习的方法，它广泛的应用于统计分类以及回归分析中。支持向量机将向量映射到一个更高维的空间里，在这个空间里建立有一个最大间隔超平面。在分开数据的超平面的两边建有两个互相平行的超平面。分隔超平面使两个平行超平面的距离最大化。假定平行超平面间的距离或差距越大，分类器的总误差越小。一个极好的指南是的《模式识别支持向量机指南》。vanderWalt 和 Barnard 将支持向量机和其他分类器进行了比拟。

4.TheApriorialgorithm

Apriori 算法是一种最有影响的挖掘布尔关联规那么频繁项集的算法。其核心是基于两阶段频集思想的递推算法。该关联规那么在分类上属于单维、单层、布尔关联规那么。在这里，所有支持度大于最小支持度的项集称为频繁项集，简称频集。

5.最大期望(EM)算法

在统计计算中，最大期望〔EM，Expectation－Maximization〕算法是在概率〔probabilistic〕模型中寻找参数最大似然估计的算法，其中概率模型依赖于无法观测的隐藏变量〔LatentVariabl〕。最大期望经常用在机器学习和计算机视觉的数据集聚〔DataClustering〕领域。

6.PageRank

PageRank 是 Google 算法的重要内容。2001 年 9 月被授予美国专利，专利人是 Google 创始人之一拉里·佩奇〔LarryPage〕。因此，PageRank 里的 page 不是指网页，而是指佩奇，即这个等级方法是以佩奇来命名的。

PageRank 根据网站的外部链接和内部链接的数量和质量俩衡量网站的价值。PageRank 背后的概念是，每个到页面的链接都是对该页面的一次投票，被链接的越多，就意味着被其他网站投票越多。这个就是所谓的“链接流行度”——衡量多少人愿意将他们的网站和你的网站挂钩。PageRank 这个概念引自学术中一篇论文的被引述的频度——即被别人引述的次数越多，一般判断这篇论文的权威性就越高。

7.AdaBoost

Adaboost 是一种迭代算法，其核心思想是针对同一个训练集训练不同的分类器(弱分类器)，然后把这些弱分类器集合起来，构成一个更强的最终分类器(强分类器)。其算法本身是通过改变数据分布来实现的，它根据每次训练集之中每个样本的分类是否正确，以及上次的总体分类的准确率，来确定每个样本的权值。将修改正权值的新数据集送给下层分类器进行训练，最后将每次训练得到的分类器最后融合起来，作为最后的决策分类器。

8.kNN:k-nearestneighborclassification

K 最近邻(k-NearestNeighbor，KNN)分类算法，是一个理论上比拟成熟的方法，也是最简单的机器学习算法之一。该方法的思路是：如果一个样本在特征空间中的 k 个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别，那么该样本也属于这个类别。

9.NaiveBayes

在众多的分类模型中，应用最为广泛的两种分类模型是决策树模型(DecisionTreeModel)和朴素贝叶斯模型〔NaiveBayesianModel，NBC〕。朴素贝叶斯模型发源于古典数学理论，有着坚实的数学根底，以及稳定的分类效率。同时，NBC 模型所需估计的参数很少，对缺失数据不太敏感，算法也比拟简单。理论上，NBC 模型与其他分类方法相比具有最小的误差率。但是实际上并非总是如此，这是因为 NBC 模型假设属性之间相互独立，这个假设在实际应用中往往是不成立的，这给 NBC 模型的正确分类带来了一定影响。在属性个数比拟多或者属性之间相关性较大时，NBC 模型的分类效率比不上决策树模型。而在属性相关性较小时，NBC 模型的性能最为良好。

10.CART:分类与回归树

CART,ClassificationandRegressionTrees。在分类树下面有两个关键的思想。第一个是关于递归地划分自变量空间的想法；第二个想法是用验证数据进行剪枝。[1]

数据挖掘常用的方法

[2]在大数据时代，数据挖掘是最关键的工作。大数据的挖掘是从海量、不完全的、有噪声的、模糊的、随机的大型数据库中发现隐含在其中有价值的、潜在有用的信息和知识的过程，也是一种决策支持过程。其主要基于人工智能，机器学习，模式学习，统计学等。通过对大数据高度自动化地分析，做出归纳性的推理，从中挖掘出潜在的模式，可以帮助企业、商家、用户调整市场政策、减少风险、理性面对市场，并做出正确的决策。目前，在很多领域尤其是在商业领域如银行、电信、电商等，数据挖掘可以解决很多问题，包括市场营销策略制定、背景分析、企业管理危机等。大数据的挖掘常用的方法有分类、回归分析、聚类、关联规那么、神经网络方法、Web 数据挖掘等。这些方法从不同的角度对数据进行挖掘。

(1)分类。分类是找出数据库中的一组数据对象的共同特点并按照分类模式将其划分为不同的类，其目的是通过分类模型，将数据库中的数据项映射到摸个给定的类别中。可以应用到涉及到应用分类、趋势预测中，如淘宝商铺将用户在一段时间内的购置情况划分成不同的类，根据情况向用户推荐关联类的商品，从而增加商铺的销售量。

(2)回归分析。回归分析反映了数据库中数据的属性值的特性，通过函数表达数据映射的关系来发现属性值之间的依赖关系。它可以应用到对数据序列的预测及相关关系的研究中去。在市场营销中，回归分析可以被应用到各个方面。如通过对本季度销售的回归分析，对下一季度的销售趋势作出预测并做出针对性的营销改变。

(3)聚类。聚类类似于分类，但与分类的目的不同，是针对数据的相似性和差异性将一组数据分为几个类别。属于同一类别的数据间的相似性很大，但不同类别之间数据的相似性很小，跨类的数据关联性很低。

(4)关联规那么。关联规那么是隐藏在数据项之间的关联或相互关系，即可以根据一个数据项的出现推导出其他数据项的出现。关联规那么的挖掘过程主要包括两个阶段：第一阶段为从海量原始数据中找出所有的高频工程组;第二阶段为从这些高频工程组产生关联规那么。关联规那么挖掘技术已经被广泛应用于金融行业企业中用以预测客户的需求,各银行在自己的 ATM 机上通过捆绑客户可能感兴趣的信息供用户了解并获取相应信息来改善自身的营销。

(5)神经网络方法。神经网络作为一种先进的人工智能技术，因其自身自行处理、分布存储和高度容错等特性非常适合处理非线性的以及那些以模糊、不完整、不严密的知识或数据为特征的处理问题，它的这一特点十分适合解决数据挖掘的问题。典型的神经网络模型主要分为三大类：第一类是以用于分类预测和模式识别的前馈式神经网络模型，其主要代表为函数型网络、感知机;第二类是用于联想记忆和优化算法的反应式神经网络模型，以 Hopfield 的离散模型和连续模型为代表。第三类是用于聚类的自组织映射方法，以 ART 模型为代表。虽然神经网络有多种模型及算法，但在特定领域的数据挖掘中使用何种模型及算法并没有统一的规那么，而且人们很难理解网络的学习及决策过程。

(6)Web 数据挖掘。Web 数据挖掘是一项综合性技术，指 Web 从文档结构和使用的集合 C 中发现隐含的模式 P，如果将 C 看做是输入，P 看做是输出，那么 Web 挖掘过程就可以看做是从输入到输出的一个映射过程。

当前越来越多的Web数据都是以数据流的形式出现的,因此对Web数据流挖掘就具有很重要的意义。目前常用的 Web 数据挖掘算法有：PageRank 算法，HITS 算法以及 LOGSOM 算法。这三种算法提到的用户都是笼统的用户，并没有区分用户的个体。目前 Web 数据挖掘面临着一些问题，包括：用户的分类问题、网站内容时效性问题，用户在页面停留时间问题，页面的链入与链出数问题等。在 Web 技术高速开展的今天，这些问题仍旧值得研究并加以解决。

编辑本段根本术语

1)数据挖掘能做以下七种不同事情

〔分析方法〕：

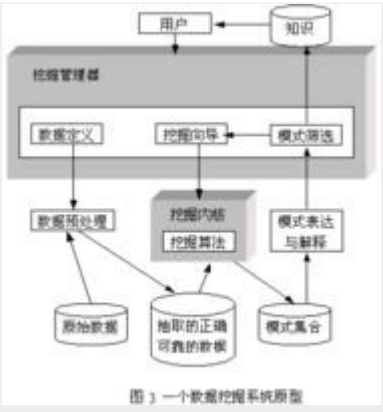


图 3 一个数据挖掘系统原型

- 分类 〔Classification〕
- 估计〔Estimation〕
- 预测〔Prediction〕
- 相关性分组或关联规那么〔Affinity grouping or association rules〕
- 聚类〔Clustering〕
- 描述和可视化〔Description and Visualization〕
- 复杂数据类型挖掘(Text, Web ,图形图像，视频，音频等)

2)数据挖掘分类

以上七种数据挖掘的分析方法可以分为两类：直接数据挖掘；间接数据挖掘

- 直接数据挖掘

目标是利用可用的数据建立一个模型，这个模型对剩余的数据，对一个特定的变量〔可以理解成数据库中表的属性，即列〕进行描述。

- 间接数据挖掘

目标中没有选出某一具体的变量，用模型进行描述；而是在所有的变量中建立起某种关系 。

- 分类、估值、预言属于直接数据挖掘；后四种属于间接数据挖掘

3)各种分析方法的简介

·分类 〔Classification〕

首先从数据中选出已经分好类的训练集，在该训练集上运用数据挖掘分类的技术，建立分类模型，对于没有分类的数据进行分类。

例子：

a. 信用卡申请者，分类为低、中、高风险

b. 故障诊断：中国宝钢集团与上海天律信息技术合作，采用数据挖掘技术对钢材生产的全流程进行质量监控和分析，构建故障地图，实时分析产品出现瑕疵的原因，有效提高了产品的优良率。

注意： 类的个数是确定的，预先定义好的

- 估计〔Estimation〕

估计与分类类似，不同之处在于，分类描述的是离散型变量的输出，而估值处理连续值的输出；分类的类别是确定数目的，估值的量是不确定的。



例子：

a. 根据购置模式，估计一个家庭的孩子个数

b. 根据购置模式，估计一个家庭的收入

c. 估计 real estate 的价值

一般来说，估值可以作为分类的前一步工作。给定一些输入数据，通过估值，得到未知的连续变量的值，然后，根据预先设定的阈值，进行分类。例如：银行对家庭贷款业务，运用估值，给各个客户记分〔Score 0~1〕。然后，根据阈值，将贷款级别分类。

- **预测〔Prediction〕**

通常，预测是通过分类或估值起作用的，也就是说，通过分类或估值得出模型，该模型用于对未知变量的预言。从这种意义上说，预言其实没有必要分为一个单独的类。预言其目的是对未来未知变量的预测，这种预测是需要时间来验证的，即必须经过一定时间后，才知道预言准确性是多少。

- 相关性分组或关联规那么〔Affinity grouping or association rules〕

决定哪些事情将一起发生。

例子：

a. 超市中客户在购置 A 的同时，经常会购置 B，即 A => B(关联规那么)

b. 客户在购置 A 后，隔一段时间，会购置 B 〔序列分析〕

- **聚类〔Clustering〕**

聚类是对记录分组，把相似的记录在一个聚集里。聚类和分类的区别是聚集不依赖于预先定义好的类，不需要训练集。

例子：

a. 一些特定病症的聚集可能预示了一个特定的疾病

b. 租 VCD 类型不相似的客户聚集，可能暗示成员属于不同的亚文化群

聚集通常作为数据挖掘的第一步。例如，"哪一种类的促销对客户响应最好？"，对于这一 类问题，首先对整个客户做聚集，将客户分组在各自的聚集里，然后对每个不同的聚集，答复以下问题，可能效果更好。

- **描述和可视化〔Description and Visualization〕**

是对数据挖掘结果的表示方式。一般只是指数据可视化工具，包含报表工具和商业智能分析产品〔BI〕的统称。譬如通过 Yonghong Z-Suite 等工具进行数据的展现，分析，钻取，将数据挖掘的分析结果更形象，深刻的展现出来。

数据挖掘十大经典算法

1。C4.5：是机器学习算法中的一种分类决策树算法，其核心算法是 ID3 算法。

2. K-means 算法：是一种聚类算法。

3.SVM：一种监督式学习的方法，广泛运用于统计分类以及回归分析中

4.Apriori ：是一种最有影响的挖掘布尔关联规那么频繁项集的算法。

5.EM：最大期望值法。

6.pagerank：是 google 算法的重要内容。

7. Adaboost:是一种迭代算法,其核心思想是针对同一个训练集训练不同的分类器然后把弱分类器集合起来，构成一个更强的最终分类器。

8.KNN:是一个理论上比拟成熟的的方法，也是最简单的机器学习方法之一。

9.Naive Bayes：在众多分类方法中，应用最广泛的有决策树模型和朴素贝叶斯〔Naive Bayes〕

10.Cart：分类与回归树，在分类树下面有两个关键的思想，第一个是关于递归地划分自变量空间的想法，第二个是用验证数据进行减枝。

编辑本段开展历程

需要是创造之母。近年来，数据挖掘引起了信息产业界的极大关注，其主要原因是存在大量数据，可以广泛使用，并且迫切需要将这些数据转换成有用的信息和知识。获取的信息和知识可以广泛用于各种应用，包括商务管理，生产控制，市场分析，工程设计和科学探索等。

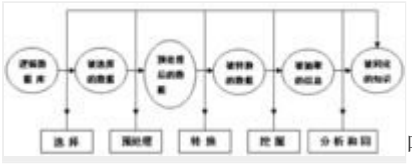
数据挖掘利用了来自如下一些领域的思想：(1) 来自统计学的抽样、估计和假设检验，(2) 人工智能、模式识别和机器学习的搜索算法、建模技术和学习理论。数据挖掘也迅速地接纳了来自其他领域的思想，这些领域包括最优化、进化计算、信息论、信号处理、可视化和信息检索。一些其他领域也起到重要的支撑作用。特别地，需要数据库系统提供有效的存储、索引和查询处理支持。源于高性能〔并行〕计算的技术在处理海量数据集方面常常是重要的。分布式技术也能帮助处理海量数据，并且当数据不能集中到一起处理时更是至关重要。

编辑本段其他资料

数据挖掘中的关联规那么

什么是关联规那么

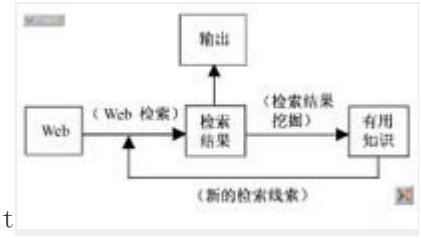
在描述有关关联规那么的一些细节之前，我们先来看一个有趣的故事： "尿布与啤酒"的故事。

在一家超市里，有一个有趣的现象：尿布和啤酒赫然摆在一起出售。但是这个奇怪的举措却使尿布和啤酒的销量双双增加了。这不是一个笑话，而是发生在美国沃尔玛连锁店超市的真实案例，并一直为商家所津津乐道。沃尔玛拥有世界上最大的数据仓库系统，为了能够准确了解顾客在其门店的购置习惯，沃尔玛对其顾客的购物行为进行购物篮分析，想知道顾客经常一起购置的商品有哪些。沃尔玛数据仓库里集中了其各门店的详细原始交易数据。在这些原始交易数据的根底上，沃尔玛利用数据挖掘方法对这些数据进行分析和挖掘。一个意外的发现是："跟尿布一起购置最多的商品竟是啤酒！经过大量实际调查和分析，揭示了一个隐藏在"尿布与啤酒"背后的美国人的一种行为模式：在美国，一些年轻的父亲下班后经常要到超市去买婴儿尿布，而他们中有30%～40%的人同时也为自己买一些啤酒。产生这一现象的原因是：美国的太太们常叮嘱她们的丈夫下班后为小孩买尿布，而丈夫们在买尿布后又随手带回了他们喜欢的啤酒。

按常规思维，尿布与啤酒风马牛不相及，假设不是借助数据挖掘技术对大量交易数据进行挖掘分析，沃尔玛是不可能发现数据内在这一有价值的规律的。

数据关联是数据库中存在的一类重要的可被发现的知识。假设两个或多个变量的取值之间存在某种规律性，就称为关联。关联可分为简单关联、时序关联、因果关联。关联分析的目的是找出数据库中隐藏的关联网。有时并不知道数据库中数据的关联函数，即使知道也是不确定的，因此关联分析生成的规那么带有可信度。关联规那么挖掘发现大量数据中项集之间有趣的关联或相关联系。Agrawal 等于 1993 年首先提出了挖掘顾客交易数据库中项集间的关联规那么问题，以后诸多的研究人员对关联规那么的挖掘问题进行了大量的研究。他们的工作包括对原有的算法进行优化，如引入随机采样、并行的思想等，以提高算法挖掘规那么的效率；对关联规那么的应用进行推广。关联规那么挖掘在数据挖掘中是一个重要的课题，最近几年已被业界所广泛研究。

关联规那么挖掘的过程

关联规那么挖掘过程主要包含两个阶段：第一阶段必须先从资料集合中找出所有的高频工程组(Frequent

Itemsets)，第二阶段再由这些高频工程组中产生关联规那么(Association Rules)。

关联规那么挖掘的第一阶段必须从原始资料集合中，找出所有高频工程组(Large Itemsets)。高频的意思是指某一工程组出现的频率相对于所有记录而言，必须到达某一水平。一工程组出现的频率称为支持度(Support)，以一个包含 A 与 B 两个工程的 2-itemset 为例，我们可以经由公式(1)求得包含{A,B}工程组的支持度，假设支持度大于等于所设定的最小支持度(Minimum Support)门槛值时，那么{A,B}称为高频工程组。一个满足最小支持度的 k-itemset，那么称为高频 k-工程组(Frequent k-itemset

)，一般表示为 Large k 或 Frequent k。算法并从 Large k 的工程组中再产生 Large k+1，直到无法再找到更长的高频工程组为止。

关联规那么挖掘的第二阶段是要产生关联规那么(Association Rules)。从高频工程组产生关联规那么，是利用前一步骤的高频 k-工程组来产生规那么，在最小信赖度(Minimum Confidence)的条件门槛下，假设一规那么所求得的信赖度满足最小信赖度，称此规那么为关联规那么。例如：经由高频 k-工程组{A,B}所产生的规那么 AB，其信赖度可经由公式(2)求得，假设信赖度大于等于最小信赖度，那么称 AB 为关联规那么。

就沃尔马案例而言，使用关联规那么挖掘技术，对交易资料库中的纪录进行资料挖掘，首先必须要设定最小支持度与最小信赖度两个门槛值，在此假设最小支持度 min_support=5% 且最小信赖度 min_confidence=70%。因此符合此该超市需求的关联规那么将必须同时满足以上两个条件。假设经过挖掘过程所找到的关联规那么「尿布，啤酒」，满足以下条件，将可接受「尿布，啤酒」的关联规那么。用公式可以描述 Support(尿布，啤酒)>=5%且 Confidence(尿布，啤酒)>=70%。其中，Support(尿布，啤酒)>=5%于此应用范例中的意义为:在所有的交易纪录资料中,至少有 5%的交易呈现尿布与啤酒这两项商品被同时购置的交易行为。Confidence(尿布，啤酒)>=70%于此应用范例中的意义为:在所有包含尿布的交易纪录资料中，至少有 70%的交易会同时购置啤酒。因此，今后假设有某消费者出现购置尿布的行为，超市将可推荐该消费者同时购置啤酒。这个商品推荐的行为那么是根据「尿布，啤酒」关联规那么，因为就该超市过去的交易纪录而言，支持了“大部份购置尿布的交易，会同时购置啤酒”的消费行为。

从上面的介绍还可以看出，关联规那么挖掘通常比拟适用与记录中的指标取离散值的情况。如果原始数据库中的指标值是取连续的数据，那么在关联规那么挖掘之前应该进行适当的数据离散化〔实际上就是将某个区间的值对应于某个值〕，数据的离散化是数据挖掘前的重要环节，离散化的过程是否合理将直接影响关联规那么的挖掘结果。

关联规那么的分类

按照不同情况，关联规那么可以进行分类如下：

1.基于规那么中处理的变量的类别，关联规那么可以分为布尔型和数值型。

布尔型关联规那么处理的值都是离散的、种类化的，它显示了这些变量之间的关系；而数值型关联规那么可以和多维关联或多层关联规那么结合起来，对数值型字段进行处理，将其进行动态的分割，或者直接对原始的数据进行处理，当然数值型关联规那么中也可以包含种类变量。例如：性别=“女”=>职业=“秘书”，是布尔型关联规那么；性别=“女”=>avg〔收入〕=2300，涉及的收入是数值类型，所以是一个数值型关联规那么。

2.基于规那么中数据的抽象层次，可以分为单层关联规那么和多层关联规那么。

在单层的关联规那么中，所有的变量都没有考虑到现实的数据是具有多个不同的层次的；而在多层的关联规那么中，对数据的多层性已经进行了充分的考虑。例如：IBM 台式机=>Sony 打印机，是一个细节数据上的单层关联规那么；台式机=>Sony 打印机，是一个较高层次和细节层次之间的多层关联规那么。

3.基于规那么中涉及到的数据的维数，关联规那么可以分为单维的和多维的。

在单维的关联规那么中，我们只涉及到数据的一个维，如用户购置的物品；而在多维的关联规那么中，要处理的数据将会涉及多个维。换成另一句话，单维关联规那么是处理单个属性中的一些关系；多维关联规那么是处理各个属性之间的某些关系。例如：啤酒=>尿布，这条规那么只涉及到用户的购置的物品；性别=“女”=>职业=“秘书”，这条规那么就涉及到两个字段的信息，是两个维上的一条关联规那么。

关联规那么挖掘的相关算法

1.Apriori 算法：使用候选项集找频繁项集

Apriori 算法是一种最有影响的挖掘布尔关联规那么频繁项集的算法。其核心是基于两阶段频集思想的递推算法。该关联规那么在分类上属于单维、单层、布尔关联规那么。在这里，所有支持度大于最小支持度的项集称为频繁项集，简称频集。

该算法的根本思想是：首先找出所有的频集，这些项集出现的频繁性至少和预定义的最小支持度一样。然后由频集产生强关联规那么，这些规那么必须满足最小支持度和最小可信度。然后使用第 1 步找到的频集产生期望的规那么，产生只包含集合的项的所有规那么，其中每一条规那么的右部只有一项，这里采用的是中规那么的定义。一旦这些规那么被生成，那么只有那些大于用户给定的最小可信度的规那么才被留下来。为了生成所有频集，使用了递推的方法。

可能产生大量的候选集,以及可能需要重复扫描数据库，是 Apriori 算法的两大缺点。

2.基于划分的算法

Savasere

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：https://d.book118.com/686143035031010105